



n|w

Swiss
Anon

I think Corolin
should also be
listed here?

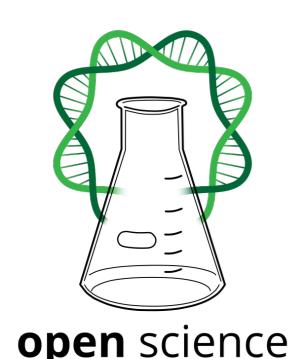


Anonymization of data for open science in psychology

1. BACKGROUND

There is a growing demand for more research data to be made openly available. The reproducibility of findings is in crisis, and more openly available data would make research more transparent and accessible. However, psychological datasets often include sensitive ~~demographic and health~~ ^{person a} information that necessitates robust privacy protection.

OPEN SCIENCE, OPEN ACCESS, OPEN DATA



OPEN ACCESS



Research data that results from publicly funded research should be:

- Findable, Accessible, Interoperable, Reusable ('FAIR principles') [1]
- therefore replicable, transparent, shareable, trustworthy, verifiable and accountable
- As open as possible, as closed as necessary

[Commission Recommendation \(EU\) 2018/790 on access to and preservation of scientific information](#)

2. METHODOLOGY

A key concern with the disclosure of personal data is whether an attacker can gain any new information about an individual.

To enable dissemination and, therefore, to open data, researchers may use methods of Statistical Disclosure Control (SDC) [2]

► SDC is the traditional approach to protecting ~~data~~ outputs against re-identification

- Non-perturbation methods (alter data without changing its actual values)
 - * Local suppression
 - * Global recoding
 - * Top and bottom coding
 - * Sampling, ...

all these methods change the values

- Perturbation methods (modify data)
 - * Noise masking *adding noise*
 - * Record swapping,
 - * Microaggregation, ...

Due to today's threats, traditional SDC methods are increasingly inadequate to protect data, necessitating advanced techniques

► Synthetic data generation

- mimics the original data
- creates artificial data that can be safely disseminated

risk can be high

(depends)

3. ILLUSTRATIVE EXAMPLE

English

As the example data were used [Holzinger and Swineford Dataset](#). The anonymization with synthetization may be performed in R package [synthpop](#) or for complex data in [simPop](#).

In our example, we evaluated the utility of synthetic data by comparing original and synthetic datasets on several metrics.

DATA UTILITY

Linking is no issue for synthetic data

Data utility refers to the usefulness of the data for the intended purpose. On the other side stands re-identification risk, which is the risk that an intruder can link a record in the released data to a specific individual in the population. So, there is a risk-utility trade-off. Balancing data utility and privacy is essential. High utility ensures synthetic data's effectiveness for research, while privacy measures minimize re-identification risk.

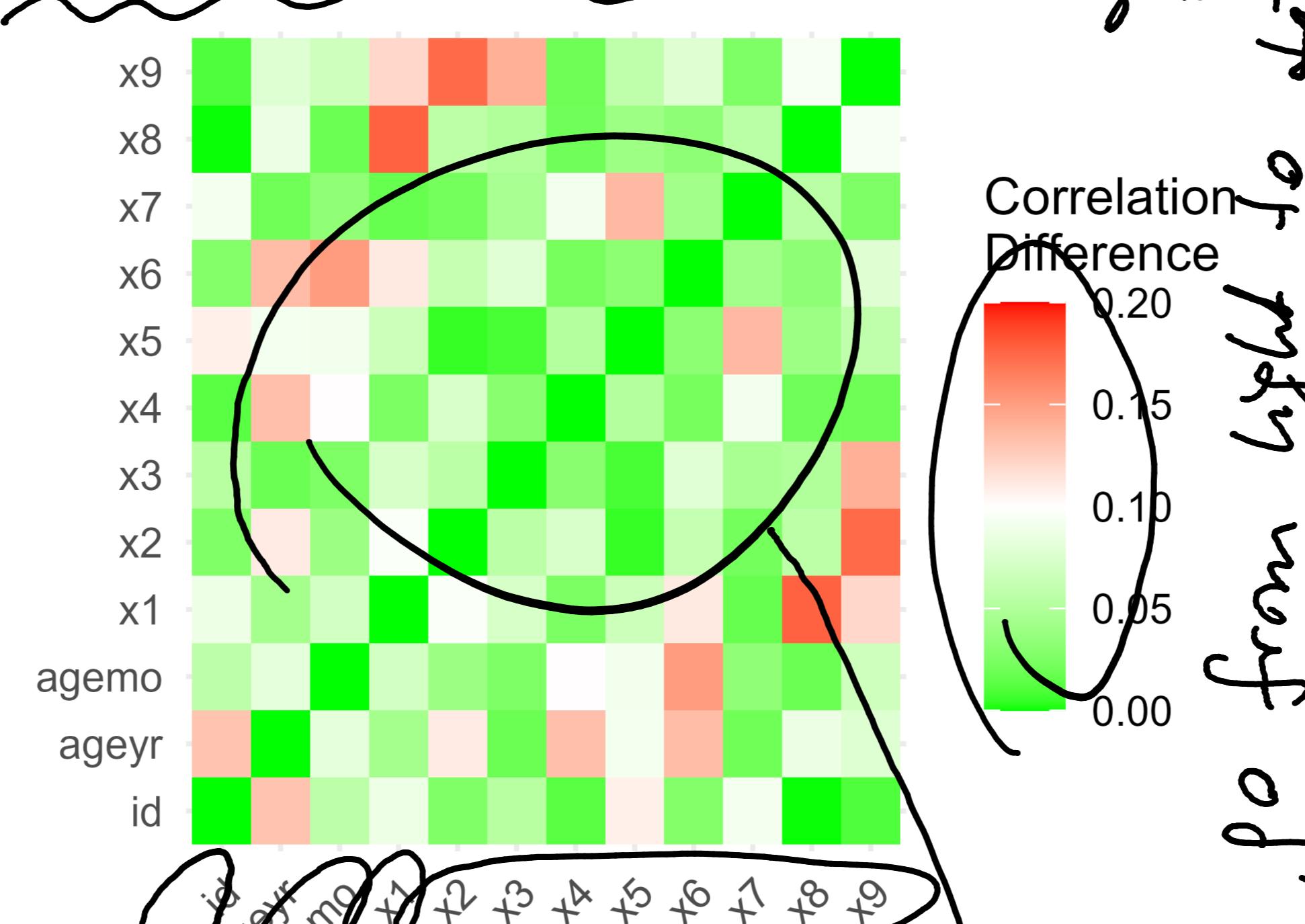


Figure 1: Difference in correlations between Original and Synthetic dataset

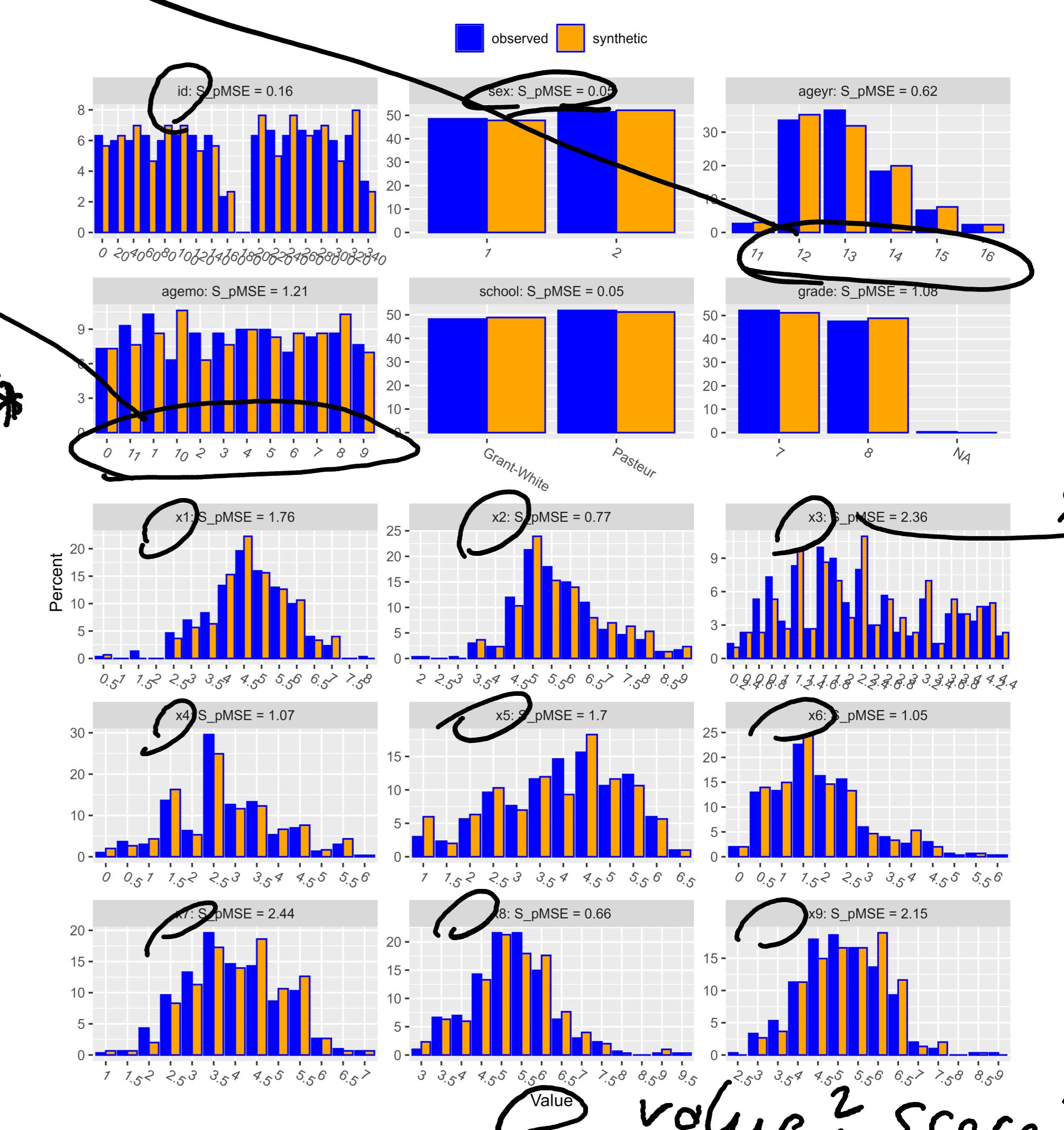


Figure 2: Difference in distributions between original and synthetic dataset

One approach to assessing the effectiveness of synthetic data generation is to compare the synthetic dataset to the original using both visual and statistical methods, focusing on distributions and correlations to ensure accurate representation.

4. Forthcoming Research

We will focus on the anonymization of longitudinal data. Given the sensitive nature of health data, we aim to develop and implement innovative tools for generating synthetic longitudinal data.

References

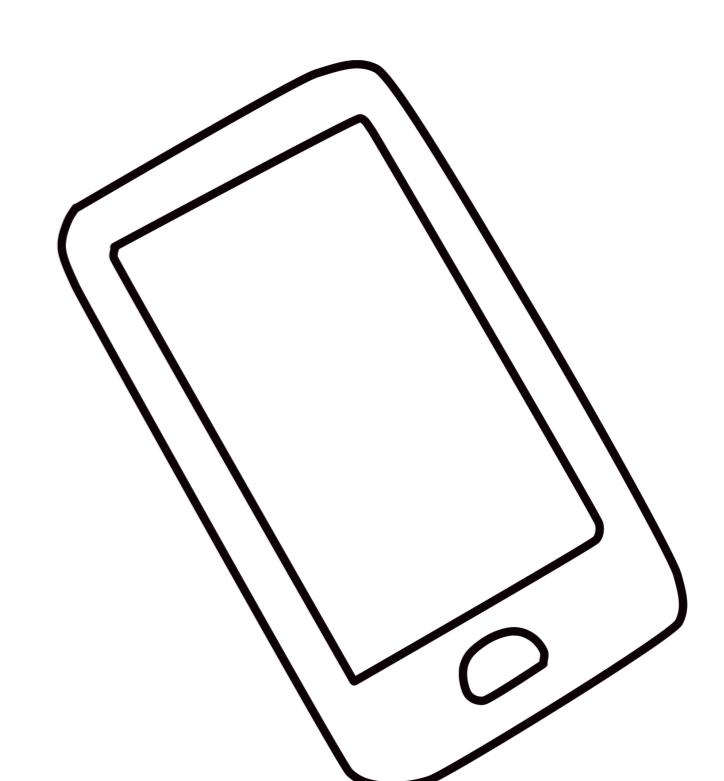
[1] European University Association. The European University Association Open Science Agenda 2025, 2022.

[2] Anco Hindepool. *Statistical disclosure control*. Wiley series in survey methodology. Wiley Chichester, West Sussex, United Kingdom, 2012.

temp. Statistical p. l... Springer 2013

Acknowledgments

This work was funded by the Swiss National Science Foundation with grant number 211751: "Harnessing event and longitudinal data in industry and health sector through privacy preserving technologies".



So bad results? I can't believe.